A distance-based method for analogical classification

Background Analogical classification has gained popularity during the last couple of decades. The main idea behind analogical classification is that we can predict the class (gender, inflection class, derivation family, etc.) of lexical items from phonological and semantic properties of those items. There are three main different approaches to analogical modelling: A(nalogical)M(odeling) (Eddington, 2002; Skousen, Lonsdale, and Parkinson, 2002; Arndt-Lappe, 2014), neural networks/deep learning (Matthews, 2010, 2013; Guzmán Naranjo, 2019) and more interpretable machine learning methods like logistic regression (Bonami and Thuilier, 2019). Although all these methods do a good job of capturing statistical properties of the data, they all suffer from different shortcomings. Neural Networks (or other ML algorithms like random forest or extreme gradient boosting trees) tend to have the best performance of all the approaches, but it is often difficult to glimps into their inner working and gain more fine-grained linguistic insight. AM offers high interpretability but the algorithm is already 30 years old, and its performance does not match that of other methods. Similarly, while AM can cope well with phonological information, it cannot handle other types of data (e.g. semantic vectors). Finally, more interpretable methods like logistic regression tend to perform poorly in cases where the structure of the predictors is non-linear, and they fail to capture more nuanced similarity relations between lexical items.

Phonological Distance-based Analogy The main insight from AM is that analogical classification (in most cases) does not follow general, rule-like patterns (e.g. all nouns ending in /k/ belong to class A), but rather work on the basis of small clusters of similar items. That is, groups of lexical items which look alike will behave alike (belong to the same class). One issue with AM is that it assumes that similarity (for the purpose of analogy) is evenly weighted across the whole word. For this reason, AM threat similarity at the end of the word identically to similarity in the middle of the word. I will show this is not consistent with what we observed in cases of analogical classification.

In this talk I will present an improved method for analogical classification: Phonological Distance-based Analogy (PDbA). PDbA works in two steps. Assuming we have N lexical items, each one belonging to one of K classes: (i) we calculate the weighted phonological distance between all lexical items we are interested in, and (ii) from the individual distances we calculate the distance of each lexical l item to the m-closest items of each class. From this set of distances we can then calculate the probability of each lexical item belonging to each class.

The idea behind a weighted phonological distance is that similarity on the right-hand edge of words is more important than similarity in the middle or left-hand side. That is, the words *pipa* is more similar to *tipa* than to *pita*, even though both pairs have the same number of identical segments. The point of calculating the distance of each lexical item to every class, is that the probability of l belonging to k_i is not only given by its phonological proximity to items of class k_i, but also to items of all other classes.

As a second step, I will show this method can be generalized as simply Distance-based Analogy, and the same principle can be applied to semantics.

Materials I will present the performance of the PDbA method on seven datasets from previous studies, shown in Table 1

Dataset	Domain	Number of classes	Number of items
Russian diminutives	derivation	3 (+ 3 rivalry)	1179
Russian inflection paradigm	inflection	108	35329
French iser-ifier	derivation	2	1413
French gender (non-suffixed nouns)	gender	2	3683
Highland Otomi verb inflection	inflection	5	1998
Spanish verb inflection	inflection	9	3034
Hausa noun inflection	inflection	16	1413

Table 1: Datasets

Additionally, I will go into the details of the Latvian noun inflection system (not previously studied from an analogical perspective).

Results Table 2 shows the performance of teh PDbA method on the different datasets, and it compares it to the performance AM, an XGBoost model and an LSTM model. The performance of PDbA is clearly at least as high or higher than the performance of other machine learning methods, and much higher than the performance of AM.

Dataset	PDbA		AM		XGBoost		LSTM	
	acc	kappa	acc	kappa	acc	kappa	acc	kappa
Ru. diminutives	0.73	0.60	0.6	0.48	0.66	0.51	0.66	0.52
Ru. inflection	0.9	0.9	0.65	0.63	0.71	0.69	0.83	0.82
Fr. iser-ifier	0.92	0.63	0.8	0.45	0.9	0.49	0.9	0.58
Fr. gender	0.83	0.64	0.76	0.59	0.82	0.63	0.81	0.62
Otomi	0.59	0.41	0.4	0.2	0.49	0.25	0.53	0.33
Sp.	0.9	0.7	0.8	0.55	0.88	0.6	0.87	0.65
Hausa	0.56	0.5	0.43	0.4	0.5	0.43	0.45	0.39

Table 2: Datasets

Concluding remarks PDbA is a new alternative to AM, with higher accuracy, and on par with modern machine learning approaches, but unlike the later, PDbA offers highly interpretable insights about the structure of the data and the way analogy operates.

References

Eddington, David (2002). “Spanish Gender Assignment in an Analogical Framework”. In: *Journal of Quantitative Linguistics* 9.1, pp. 49–75.

Skousen, Royal, Deryle Lonsdale, and Dilworth B. Parkinson (2002). *Analogical Modeling: An Exemplar-Based Approach to Language*. Amsterdam: John Benjamins.

Matthews, Clive A. (2010). “On the Nature of Phonological Cues in the Acquisition of French Gender Categories: Evidence from Instance-Based Learning Models”. In: *Lingua* 120.4, pp. 879–900.

— (Sept. 2013). “On the Analogical Modelling of the English Past-Tense: A Critical Assessment”. In: *Lingua* 133, pp. 360–373.

Arndt-Lappe, Sabine (2014). “Analogy in Suffix Rivalry: The Case of English *-Ity* and *-Ness*”. In: *English Language and Linguistics* 18.3, pp. 497–548.

Bonami, Olivier and Juliette Thuilier (2019). “A Statistical Approach to Rivalry in Lexeme Formation: French-Iser and-Ifier”. In: *Word Structure* 12.1, pp. 4–41.

Guzmán Naranjo, Matías (2019). *Analogical Classification in Formal Grammar*. Empirically Oriented Theoretical Morphology and Syntax. Language Science Press.